2012 IEEE 36th International Conference on Computer Software and Applications

Disease-disease relationships for rheumatic diseases

Web-based biomedical textmining and knowledge discovery to assist medical decision making

Andreas Holzinger & Klaus-Martin Simonic
Institute for Medical Informatics, Statistics & Documentation
Medical University Graz
Graz, Austria
{andreas.holzinger, klaus.simonic}@medunigraz.at

Pinar Yildirim
Department of Computer Engineering
Faculty of Engineering and Architecture
Okan University
Istanbul, Turkey
pinar.yildirim@okan.edu.tr

***Abstract*—The MEDLINE database (Medical Literature Analysis and Retrieval System Online) contains an enormously increasing volume of biomedical articles. There is urgent need for techniques which enable the discovery, the extraction, the integration and the use of hidden knowledge in those articles. Text mining aims at developing technologies to help cope with the interpretation of these large volumes of publications. Co-occurrence analysis is a technique applied in text mining and the methodologies and statistical models are used to evaluate the significance of the relationship between entities such as disease names, drug names, and keywords in titles, abstracts or even entire publications. In this paper we present a method and an evaluation on knowledge discovery of disease-disease relationships for rheumatic diseases. This has huge medical relevance, since rheumatic diseases affect hundreds of millions of people worldwide and lead to substantial loss of functioning and mobility. In this study, we interviewed medical experts and searched the ACR (American College of Rheumatology) web site in order to select the most observed rheumatic diseases to explore disease-disease relationships. We used a web based text-mining tool to find disease names and their co-occurrence frequencies in MEDLINE articles for each disease. After finding disease names and frequencies, we normalized the names by interviewing medical experts and by utilizing biomedical resources. Frequencies are normally a good indicator of the relevance of a concept but they tend to overestimate the importance of common concepts. We also used Pointwise Mutual Information (PMI) measure to discover the strength of a relationship. PMI provides an indication of how more often the query and concept co-occur than expected by change. After finding PMI values for each disease, we ranked these values and frequencies together. The results reveal hidden knowledge in articles regarding rheumatic diseases indexed by MEDLINE, thereby exposing relationships that can provide important additional information for medical experts and researchers for medical decision-making.**

Keywords—Biomedical text mining, rheumatic diseases, disease–disease relationships, co-occurrence analysis, Pointwise Mutual Information (PMI)

I. INTRODUCTION

The MEDLINE database is the primary resource for biomedical researchers. A wealth of scientific information is available in this database and provides knowledge on relationships between biomedical concepts including genes, diseases, and cellular processes.

A commonly used method to establish such relationships between biomedical concepts from literature is co-occurrence. Apart from its use in knowledge retrieval, the co-occurrence method is also well suited to discovering new, hidden relationships between biomedical concepts [1].

MEDLINE provides scientific information of more than 20 million articles that have been published in biomedical journals. However, all the information contained in the database is stored as text. The rapid growth of these collections makes it increasingly difficult for humans to access the required data in a convenient and effective manner. In order to make this data accessible, usable and useful, smart information retrieval systems that can operate on these non-standardized entries (often sloppy called: “free text”) are essential [2]. Consequently, there is a strong necessity of developing methods for automatic extraction of relevant information (such as keywords related with diseases) from the literature, which is written in natural language [3].

The aim of this study is to explore disease-disease relationships for rheumatic diseases. Rheumatic diseases include the pathogenesis, diagnosis, and management of over 100 complex and scientifically interesting diseases, including autoimmune diseases, arthritis, and musculo-skeletal conditions. Rheumatologists care for a wide array of patients from children to senior citizens. Rheumatic diseases affect hundreds of millions of people worldwide and lead to substantial loss of physical function and mobility. For example, rheumatoid arthritis is an inflammatory systemic disease that predominantly affects the joints. It is the most common form of arthritis, and has considerable social implications due to the costs incurred and the loss of productivity caused by the disease

0730-3157/12 $26.00 © 2012 IEEE
DOI 10.1109/COMPSAC.2012.77

IEEE computer society

progression that may culminate in occupational disability. Depending on the stage of the disease, the destructive process often also affects the surrounding connective tissue leading to deformation and functional disorders of the affected organs. The clinical picture of the disease varies with respect to the number and pattern of afflicted joints observed, possible involvement of internal organs, and the course of the disease [4]. In addition, the pathogenesis of many rheumatic diseases remains incompletely understood and is associated with higher mortality. Considering this high importance of rheumatic diseases we developed a study to find some undiscovered co-occurrence based disease-disease relationships from MEDLINE.

II. Related Work

There are several studies aimed at knowledge discovery for rheumatic diseases in literature. In a similar study, Yildirim et al. introduced a method for extracting hidden patterns in rheumatic diseases by using the Medline database. They found symptoms in related articles and then created a dataset with the frequencies of symptoms for each disease and applied hierarchical clustering analysis to find similarities between the diseases [5]. Giacomozzi et al. investigated the potential use of Peak Pressure Curves (PPC) and Normalized Vertical Force Curves (NVFC) to classify patients with rheumatic arthritis and applied cluster analysis to patient data [6]. Liu et al. tested the ability of peripheral blood gene expression profiles to predict future disease severity in patients with early rheumatoid arthritis and utilized unsupervised and supervised algorithms to identify “predictor genes” whose combined expression levels correlated with follow-up disease severity scores [7]. Cooper et al. carried out a study of patients with osteo-arthritis based on epidemiological data and used logistic regression to test for overall clustering of osteoarthritis between joint sites [8]. In a work related with expert comments on magnetic resonance images (MRI) diagnoses calculations of significant co-occurrences of diseases and defined regions of the human body, in order to identify possible risks for health has been performed [9].

III. Method

The steps we applied in our study to discover hidden patterns for rheumatic diseases are as follows:

- The most common eight rheumatic diseases were selected by interviewing a medical expert and searching the American College of Rheumatology website. Table I shows the number of articles in Medline for selected rheumatic diseases.
- Diseases and co-occurrence frequencies were extracted and Pointwise Mutual Information (PMI) values were calculated for each rheumatic disease by using a biomedical text-mining tool.
- Disease names have some variations such as synonyms. These names were normalized to one specific name. For example, Wegener's Granulomatosis and Wegener's Granuloma indicate the same meaning and can be mapped to Wegener's Granulomatosis.

There are several web-based tools for analyzing the articles. Most of them provide the analysis of co-occurrence between entities. In this study, the FACTA text mining tool was used to extract diseases from the relevant articles. FACTA is developed by the National Centre of Text Mining (NACTEM). It is a text search engine for MEDLINE abstracts designed particularly to help users browse biomedical concepts (e.g. genes/proteins, diseases, enzymes and chemical compounds). The distinct advantage of FACTA is that it delivers real-time responses whilst being able to accept flexible queries.

FACTA covers six categories of biomedical concepts: human genes/proteins, diseases, symptoms, drugs, enzymes and chemical compounds. The concepts appearing in the documents are recognized by dictionary matching. The Unified Medical Language System (UMLS) is used for the normalization of terms for diseases and symptoms. UMLS constitutes a valuable lexical resource integrating a thesaurus and multilingual vocabulary database of health-related concepts as well as the semantic relationships between them. FACTA receives a query from the user as input. A query can be a concept name such as “rheumatoid arthritis“, a concept ID or a combination of the two. The system then retrieves all the documents that match the query from MEDLINE using word/concept indexes. The concepts contained in the documents are then counted and ranked according to their relevance to the query. For the input query “rheumatoid arthritis”, with disease as a selected concept, the system retrieves 94834 documents from MEDLINE. The results are displayed as a table and ranked by their frequencies, which indicate how many times the selected concept appears in the articles. For example, “polyarthritis” appears 4393 times with “rheumatoid arthritis” [10].

TABLE I. Number of Articles for the Selected Rheumatic Diseases in MEDLINE

Selected Rheumatic Diseases	*Number of Articles*
Rheumatoid arthritis	94834
Fibromyalgia	5907

Selected Rheumatic Diseases	*Number of Articles*
Gout	10438
Henoch-Schoenlein purpura	3276
Osteoarthritis	41613
Polyarteritis nodosa	5795
Systemic sclerosis	10006
Wegener's Granulomatosis	4769

Statistical techniques play an important role for text mining studies. There are several measures of co-occurrence analysis:

The simplest method of identifying relationships is using the co-occurrence assumption: terms that appear in the same texts tend to be related. For example, if a protein is mentioned often in the same abstracts as a disease, it is reasonable to hypothesize that this protein is involved in some aspect to this disease. The degree of co-occurrence can be quantified statistically to rank and eliminate statistically weak co-occurrences [10].

Pointwise Mutual Information is an ideal measure of word association norms based on information theory and we selected this measure to analyze rheumatic diseases. PMI compares the probability of observing two items together with the probabilities of observing two items independently. Therefore, it can be used to estimate whether the two items have a genuine association or are observed at random [12].

Let two words, w_i and w_j, have probabilities $P(w_i)$ and $P(w_j)$. Their mutual information $PMI(w_i, w_j)$ is defined as:

$$PMI(w_i, w_j) = \log\left(\frac{P(w_i, w_j)}{P(w_i)\,P(w_j)}\right)$$

For w_i denoting *rheumatoid arthritis* and w_j representing *diffuse scleritis* the following simple calculation yields:

$$P(w_i) = \frac{94{,}834}{20{,}033{,}079}, \quad P(w_j) = \frac{74}{20{,}033{,}079}$$

$$P(w_i, w_j) = \frac{13}{94{,}834}, \quad PMI(w_i, w_j) = 7{,}7.$$

TABLE II.

TABLE III. CO-OCCURRENCE DATA

	Count
Total articles in Medline	20.033.079
Rheumatoid arthritisand diffuse scleritis co-occurrence frequency	13
Rheumatoid arthritis individual frequency	94.834
Diffuse scleritis individual frequency	74

IV. RESULTS

In our study, co-occurrence frequency and PMI values were calculated for each disease and the first ten diseases having higher values were ranked. Table III shows ranked diseases according to both of measures and Figure 1-16 show the graphical representations of the results. These results reveal that new relationships can be found based on most frequent diseases related a specific rheumatic disease.

TABLE IV. RANKED DISEASES BY FREQUENCY AND PMI

RHEUMATOID ARTHRITIS

Rank	*Disease*	*Frequency*
1	Arthritis	15191
2	Osteoarthritis	8045
3	Arthritis, Juvenile Rheumatoid	7982
4	Systemic lupus erythematosus	7176
5	Polyarthritis	4393
6	Autoimmune Diseases	3993
7	Ankylosing spondylitis	3469
8	Synovitis	3264
9	Joint Diseases	3139
10	Tumor	2213

Rank	*Disease*	*PMI*
1	Diffuse scleritis	7,7
2	Subacute arthritis	7,7
3	Juvenile spondyloarthropathy	7,7
4	Mitral stenosis and aortic regurgitation	7,7
5	Seropositive rheumatoid arthritis	7,7
6	Monoarticular Arthritis	7,7
7	Bursopathy	7,7
8	Monarticular juvenile rheumatoid arthritis	7,7
9	Secondary fibrositis	7,7
10	Seronegative rheumatoid arthritis	7,7

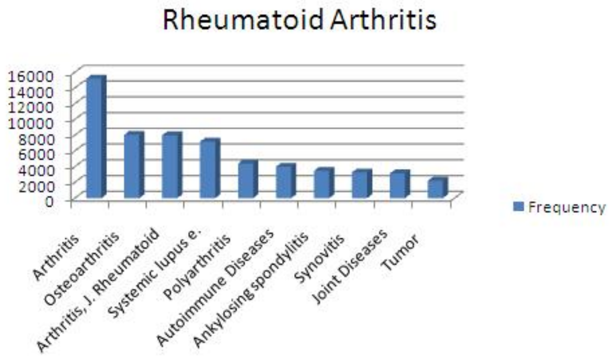

Figure 1. Rheumatoid Arthritis and diseases by frequency.

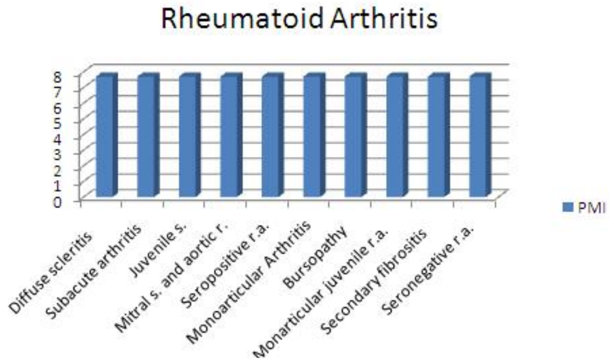

Figure 2. Rheumatoid Arthritis and diseases by PMI.

FIBROMYALGIA

Rank	*Disease*	*Frequency*
1	Depression	845
2	Chronic pain	627
3	Rheumatoid arthritis	600
4	Chronic fatigue syndrome	527
5	Osteoarthritis	332
6	Arthritis	300
7	Sleep Disorders	259
8	Irritable bowel syndrome	244
9	Stress, Psychological	204
10	Systemic lupus erythematosus	202

Rank	*Disease*	*PMI*
1	Painful functional bowel disorder	11,7
2	Primary Fibromyalgia	11,7
3	Secondary fibrositis	11,7
	Primary fibrositis	11,3
5	Myofibrositis	11
6	Sexual pain disorders	9,7
7	Interstitial myositis	9,7
8	Spurious diarrhoea	9,7
9	Myofascial pain	9,7
10	Stress related disorder	9,4

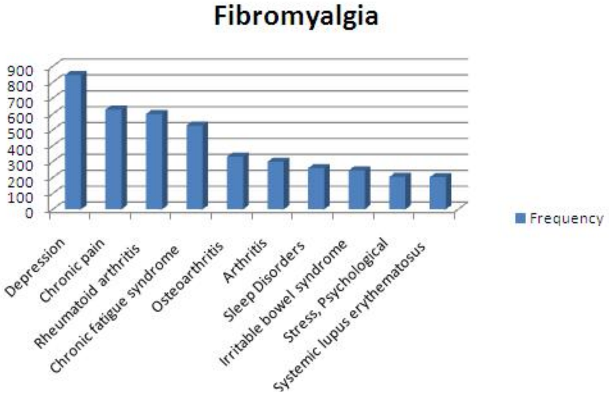

Figure 3. Fibromyalgia and diseases by frequency.

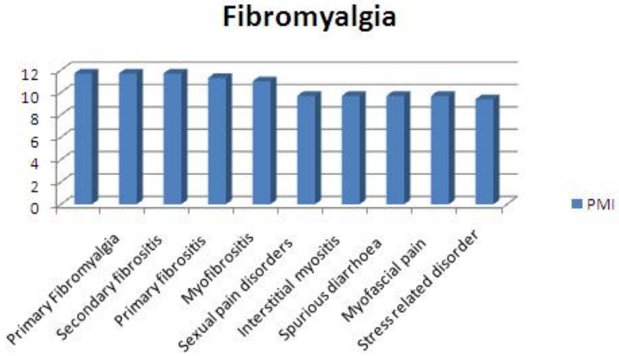

Figure 4. Fibromyalgia and diseases by PMI.

GOUT

Rank	*Disease*	*Frequency*
1	Arthritis	1661
2	Hyperuricemia	1334
3	Rheumatoid arthritis	1280
4	Hypertension	739
5	Osteoarthritis	730
6	Kidney Diseases	707
7	Diabetes	635
8	Joint Diseases	563
9	Rheumatic Diseases	529
10	Arthritis, Gouty	496

Rank	*Disease*	*PMI*
1	Secondary gout	10,9
2	Acquired reactive perforating dermatosis	10,9
3	Saturnine gout	10,9
4	Chronic gouty arthritis	10,9
5	Visceral gout	10,9
6	Gout and other crystal arthropathies	10,9
7	Drug-induced gout	10,9
8	Chronic tophaceous gout	10,9
9	Chronic gouty nephropathy	10,9
10	Acute lead nephropathy	10,9

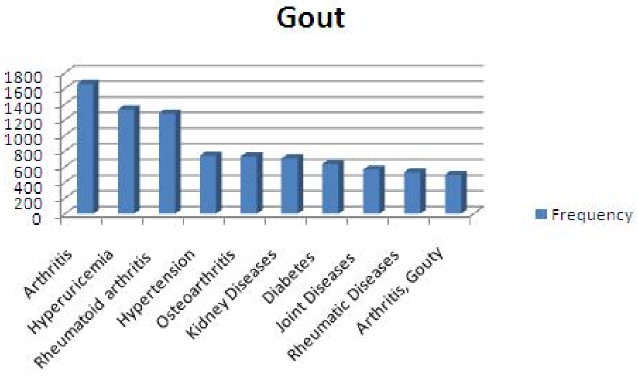

Figure 5. Gout and diseases by PMI.

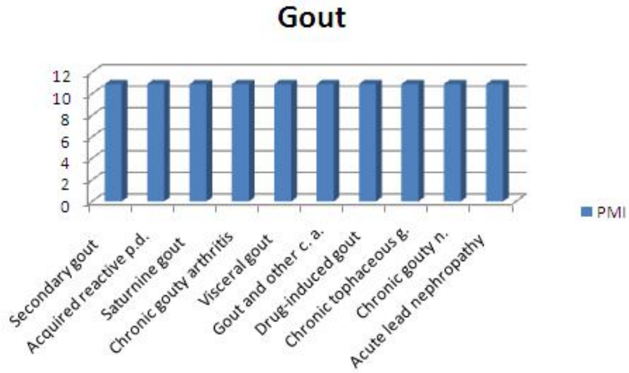

Figure 6. Gout and diseases by PMI.

HENOCH SCHOENLEIN PURPURA

Rank	*Disease*	*Frequency*
1	Vasculitis	736
2	Nephritis	458
3	Glomerulonephritis	416
4	Kidney Diseases	336
5	IgA nephropathy	288
6	Proteinuria	257
7	Rash	183
8	Systemic lupus erythematosus	134
9	Nephrotic syndrome	131
10	Polyarteritis Nodosa	121

Rank	*Disease*	*PMI*
1	Sindbis fever	12,6
2	Empedobacter brevis infection	12,6
3	Henoch-Schoenlein nephritis	12,3
4	Dysproteinemic purpura	12
5	Rheumatic purpura	12
6	Purpura rheumatica	11,9
7	Squamous cell carcinoma of bronchus	11,6
8	Homozygous Factor V Leiden mutation	11,6
9	Infantile acute hemorrhagic edema	11,4
10	Purpura fulminans	11

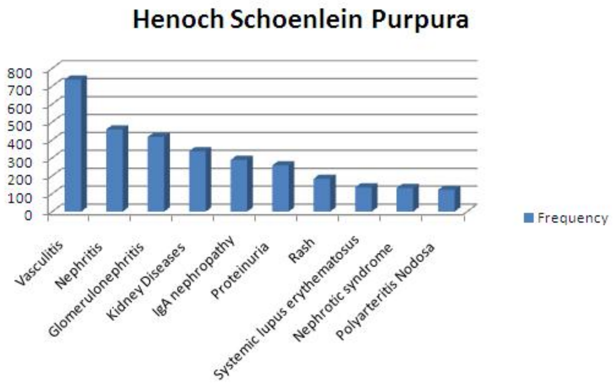

Figure 7. Henoch Schoenlein Purpura and diseases by frequency.

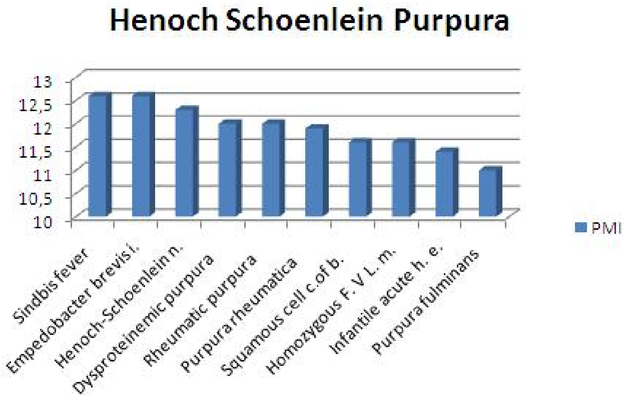

Figure 8. Henoch Schoenlein Purpura and diseases by PMI.

OSTEOARTHRITIS

Rank	*Disease*	*Frequency*
1	Rheumatoid arthritis	7953
2	Osteoarthritis, Knee	7218
3	Osteoarthritis, Hip	5173
4	Arthritis	5132
5	Joint Diseases	3560
6	Blind	1328
7	Arthrodesis	1201
8	Synovitis	1074
9	Osteoporosis	979
10	Obesity	927

Rank	*Disease*	*PMI*
1	Nodal osteoarthritis	8,9
2	Acquired Deafness	8,9
3	Osteoarthritis of the shoulder	8,9
4	Osteoarthritis aggravated	8,9
5	Head deformity	8,9
6	Finger osteoarthritis	8,9
7	EDM5	8,9
8	Osteoarthritis of the wrist	8,9
9	Major osseous defects	8,9
10	Osteoarthritis of the lumbar spine	8,9

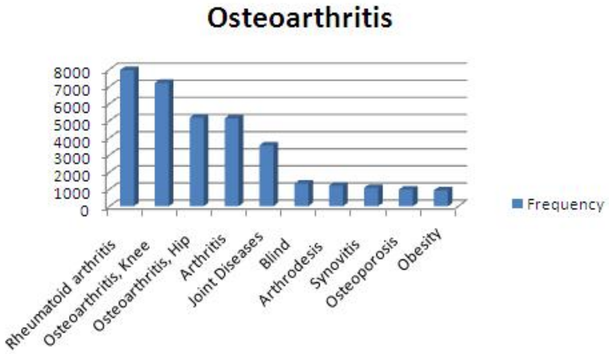

Figure 9. Osteoarthritis and diseases by frequency.

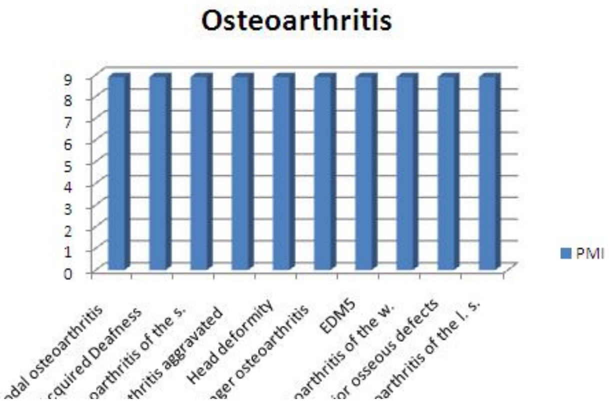

Figure 10. Osteoarthritis and diseases by PMI.

SYSTEMIC SCLEROSIS

Rank	*Disease*	*Frequency*
1	Scleroderma, Systemic	7834
2	Scleroderma	2138
3	Systemic lupus erythematosus	1967
4	Multiple sclerosis	1438
5	Rheumatoid arthritis	1181
6	Autoimmune Diseases	957
7	Raynaud Disease	818
8	Raynaud's phenomenon	791
9	Connective Tissue Diseases	635
10	Pulmonary Fibrosis	597

Rank	Disease	PMI
1	Systemic sclerosis sine scleroderma	11
2	Occupational scleroderma	11
3	Alezzandrini syndrome	11
4	Sclerosis of the skin	11
5	Enteroparesis	11
6	Lung involvement in systemic sclerosis	11
7	Scleroderma, Limited	10,4
8	Diffuse scleroderma	10,2
9	Scleroderma renal crisis	10,1
10	Nuclear sclerotic cataract	10

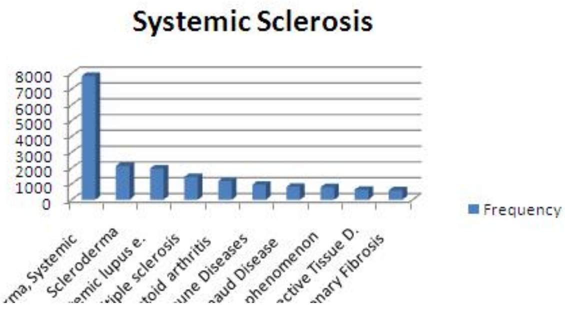

Figure 11. Systemic Sclerosis and diseases by frequency.

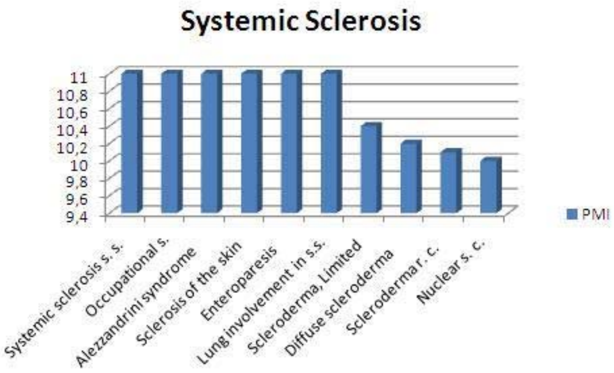

Figure 12. Systemic Sclerosis and diseases by PMI.

POLYARTERITIS NODOSA

Rank	*Disease*	*Frequency*
1	Vasculitis	1569
2	Systemic lupus erythematosus	719
3	Wegener's granulomatosis	680
4	Rheumatoid arthritis	487
5	Hepatitis B	414
6	Glomerulonephritis	411
7	Scleroderma, Systemic	333
8	Kidney Diseases	303
9	Churg-Strauss syndrome	284
10	Collagen Diseases	281

Rank	*Disease*	*PMI*
1	Cheshire cat syndrome	11,8
2	Silicoarthritis	11,8
3	Benign cutaneous periarteritis nodosa	11,8
4	Cutaneous polyarteritis nodosa	11,8
5	Juvenile polyarteritis	11,3
6	Polyangiitis overlap syndrome	11,1
7	Necrotizing angiitis	11
8	Periarteritis	10,9
9	Small intestine gangrene	10,8
10	Buccal ulceration	10,8

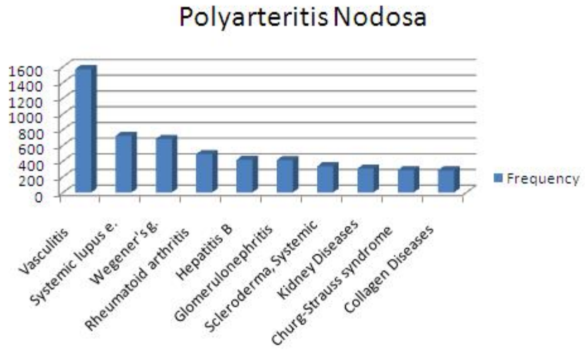

Figure 13. Polyarteritis Nodosa and diseases by frequency.

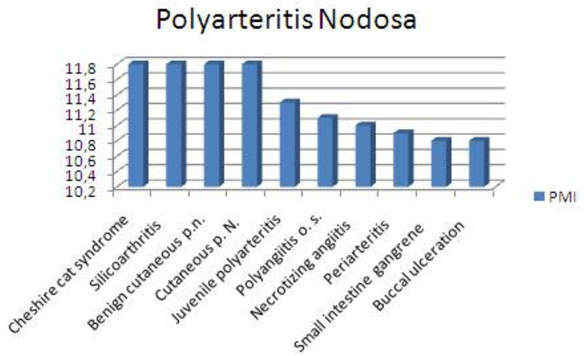

Figure 14. Polyarteritis Nodosa and diseases by PMI.

WEGENER'S GRANULOMATOSIS

Rank	*Disease*	*Frequency*
1	Vasculitis	1654
2	Glomerulonephritis	534
3	Polyarteritis Nodosa	454
4	Churg-Strauss syndrome	383
5	Systemic lupus erythematosus	335
6	Lung Diseases	311
7	Kidney Diseases	235
8	Rheumatoid arthritis	225
9	Autoimmune Diseases	207
10	Granuloma	203

Rank	*Disease*	*PMI*
1	Intranasal adhesions	12
2	Fulminant enterocolitis	12
3	Refractory Wegener's granulomatosis	12
4	Nasal septal necrosis	12
5	Trabeculitis	11
6	damage; renal	11
7	Histiocytic proliferation	11
8	Sclerosis of the skin	11
9	Adrenal infarction	11
10	Focal vasculitis	11

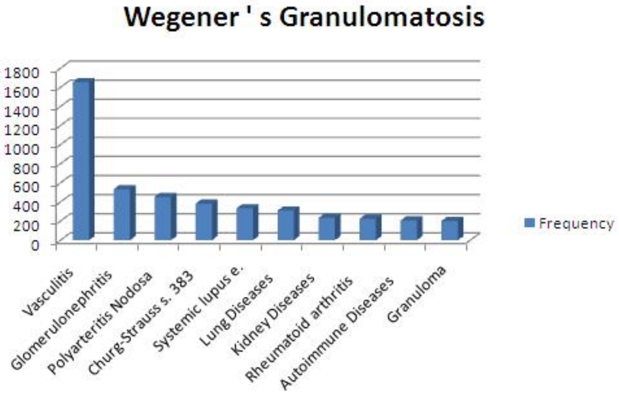

Figure 15. Wegener's Granulomatosis and diseases by frequency.

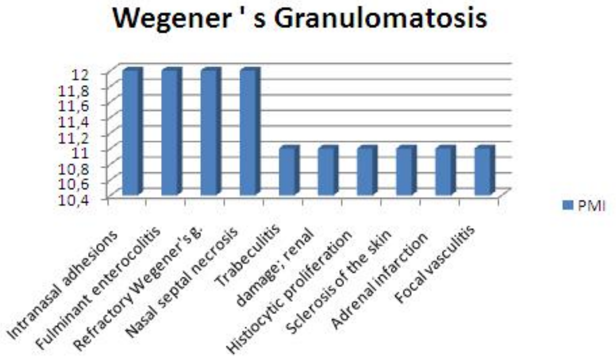

Figure 16. Wegener's Granulomatosis and diseases by PMI.

V. Discussion and Conclusion

Biomedical research aims to discover new medical knowledge to support better diagnosis, decision making and treatment. Biomedical literature, such as indexed by the MEDLINE database, provides many opportunities to reach this goal. In this work we focused on knowledge discovery for rheumatic diseases and utilized a web-based biomedical text mining system to obtain some measurements. Rheumatic diseases can lead to several harmful effects in the body and organs, and research on these diseases is an important part of physical medicine. The diseases can be related to other diseases. We considered this problem and investigated pertinent knowledge in relevant articles to gather some relationships between rheumatic diseases and other diseases. Apart from biomedical text mining techniques, we applied statistical co-occurrence analysis and selected co-occurrence frequency along with pointwise mutual information (PMI) measures. Our results show that a lot of diseases can have relationships with rheumatic diseases. Medical experts can interpret the links between these diseases; hence develop ideas to discover new knowledge from this information. However, some relationships may only be statistically valid, therefore medical experts have to evaluate and observe these results through specific clinical studies.

Further work includes expanding our steps by using a medical dictionary, such as SNOMED-CT, to gather not only disease names, but also diagnosis and treatment names/codes to see whether and to what extent the identification of diseases in articles is increased by checking for diagnosis and treatment of the diseases. Moreover, since mutual information is a measure of the information overlap between two random variables, we also plan to experiment with measuring entropy in a further study.

Acknowledgment

We cordially thank Dr. Cinar Ceken for her help and medical expert advice. We also thank the COMPSAC reviewers for their thorough review and helpful comments to further improve our paper.

REFERENCES

[1] R. Frijters, M. van Vugt, R. Smeets, R. van Schaik, J. de Vlieg and W. Alkema, “Literature Mining for the Discovery of Hidden Connections between Drugs, Genes and Diseases”,PLoS Comput Biol. vol.0 6, no. 9,p. e1000943, 2010.

[2] M. Kreuzthaler, M. D. Bloice, L. Faulstich, K. M. Simonic, and A. Holzinger, "A Comparison of Different Retrieval Strategies Working on Medical Free Texts", Journal of Universal Computer Science, vol. 17, no. 7, pp. 1109–1133, 2011..

[3] Y. Liua, M. Brandona, S. Navathea, R. Dingledine , and B. J. Ciliax,”Text Mining Functional Keywords Associated with Gene”, MEDINFO M. Fieschi et al. (Eds) Amsterdam: IOS Press 2004..

[4] K. M. Simonic, A. Holzinger, M. Bloice, and J. Hermann,” Optimizing Long-Term Treatment of Rheumatoid Arthritis with Systematic Documentation”,5th International Conference on Pervasive Computing Technologies for Healthcare, Dublin: IEEE, pp. 550–554, 2011.

[5] P.Yildirim, Ç. Çeken, R. Hassanpour, M. R. Tolun, ”Prediction of similarities between rheumatic diseases”, Journal of Medical Systems. (in print)

[6] C. Giacomozzi,F. Martelli,A. Nagel, A. Schmiegel, D. Rosenbaum, ” Cluster Analysis to Classify Gait Alterations in Rheumatoid Arthritis Using Peak Pressure Curves”, Gait&Posture, vol. 29, pp. 220–224, 2009.

[7] Z.Liu, T. Sokka, K.Maas, N.J. Olsen, and T.M. Aune, “Prediction of Disease Severity in Patients with Early Rheumatoid Arthritis by Gene Expression Profiling”, Human Genomics and Proteomics, vol. 2009.

[8] C. Cooper, P. Egger,D. Coggon,D.J. Hart,T. Masud,F. Cicuttini, D.V. Doyle and T.D. Spector, “Generalized Osteoarthritis in Women: Pattern of Joint Involvement and Approaches to Definition for Epidemiological Studies”, Journal of Rheumatology, vol. 23, no. 11, pp. 1938–1942, 1996.

[9] A. Holzinger, R. Geierhofer, F. Modritscher, and R. Tatzl, "Semantic Information in Medical Information Systems: Utilization of Text Mining Techniques to Analyze Medical Diagnoses," *Journal of Universal Computer Science,* vol. 14, pp. 3781-3795, 2008.

[10] M. Krallinger, F. Leither and A. Valencia, “Analysis of Biological Processes and Diseases Using Text Mining Approaches”, Methods in Molecular Biology, vol. 593, pp. 341–382, 2010.

[11] Y.Tsuruoka, J. Tsujiiand S. Ananiadou,”FACTA: a text search engine for finding associated biomedicalconcepts”, Bioinformatics, vol. 24, no. 21, pp. 2559–256, 2008.

[12] R.Rodriguez-Esteban, “Biomedical Text Mining and Its Applications”, PloS Computational Biology, vol.5, no.12, p. e1000597, 2009.